

Interval estimations in metrology

G Mana¹ and C Palmisano²

¹INRIM - Istituto Nazionale di Ricerca Metrologica, Str. delle Cacce 91, 10135 Torino, Italy

²UNITO - Università di Torino, Dipartimento di Fisica, v. P. Giuria, 1 10125 Torino, Italy

E-mail: g.mana@inrim.it

Abstract. This paper investigates interval estimation for a measurand that is known to be positive. Both the Neyman and Bayesian procedures are considered and the difference between the two, not always perceived, is discussed in detail. A solution is proposed to a paradox originated by the frequentist assessment of the long-run success rate of Bayesian intervals.

Submitted to: *Metrologia*

PACS numbers: 02.50.Cw, 02.50.Tt, 06.20.Dk, 07.05.Kf

1. Introduction

The Neyman and Bayes viewpoints about how to carry out interval estimation [1, 2, 3, 4] lead to different uncertainty statements. Since they are calculating different intervals, there is a debate over the meaning of confidence level and coverage probability for an uncertainty statement. A non-exhaustive list of papers investigating this issue is [5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. A simple problem that makes the viewpoints' differences evident is when there is a measurement of a real quantity that is small with respect to the measurement uncertainty and that it is known to be have a well defined sign [18, 19]. In [20, 21] Willink showed that, in a Monte Carlo simulation of repeated Gaussian measurements of the same measurand, the long-run success rate of Bayesian intervals to encompass the measurand disagrees with the expected value. This paper presents the results of an investigation designed to understand the basic concepts of interval estimation and to explain this paradoxical result.

When reporting the uncertainty of measurements, the awareness of the differences between the Neyman and Bayesian approaches is essential. Interval estimation is a procedure to find a pair of values that succeeds in including the measurand with a stipulated probability. In the Neyman approach, the focus is on different interval estimations, given the same measurand value. In the Bayesian approach, the focus is on different measurand values, given the same interval. After reviewing the Neyman and Bayesian solutions, the paper illustrates interval estimation given a Gaussian sample of a positive quantity. Eventually, it shows that both approaches achieve the stipulated success rate.

2. Interval estimation

Before the measurement is carried out, the measured value, x , can be viewed as an unknown member of a population described by a probability distribution, $P_x(\xi|a)$, parameterised by the measurand value a , which – though unknown – has a fixed value. In $P_x(\xi|a)$, the letter ξ is a dummy variable which labels the space of the possible x values; the vertical bar indicates that the probability density in $x = \xi$ is conditioned on a measurand value equal to a .

The probability distribution of the measurand values enters our considerations because we face a range of possible values, but we are not able to figure out what it is. Probability assignments to the a values and probability calculus make our knowledge quantitative and allow us to come to sensible decisions. Prior to the measurement, the measurand value can be viewed as an unknown member of a population described by the probability distribution $\pi(\phi)$ – where ϕ labels the possible a values. When the measured value x_0 is on hand, $\pi(\phi)$ must be updated to $P_a(\phi|x_0)$, where x_0 is a known parameter. These distributions are linked by the Bayes theorem

$$P_a(\phi|x_0) = \frac{L(\phi; x_0)\pi(\phi)}{Z(x_0)}, \quad (1)$$

where $L(\phi; x_0) = P_x(x_0|\phi)$ is the likelihood function and $Z(x_0)$ is a normalising factor.

Given the measurement result x_0 , interval estimation is the problem of finding an interval – $[a_1, a_2]$, which is called a credible (or coverage) interval – such that the measurand value in it with a predetermined probability – $\text{Prob}(a \in [a_1, a_2]|x_0)$, which is called coverage probability and is represented by α . Therefore, the interval end-points are the solutions of

$$\text{Prob}(a \in [a_1, a_2]|x_0) = \alpha. \quad (2)$$

It must be noted that (2) is conditioned on the fixed measured value x_0 .

2.1. Neyman: confidence intervals

According to Neyman, it is meaningless to assign probabilities to the possible measurand value; he discarded (1) and (2) and proposed a statistic – that is, a function of the measurement result – having, in a long series of repeated application to different measured values of the same measurand, a success rate of including the measurand equal to the coverage probability [2]. The Neyman interval, $[\underline{a}, \bar{a}]$, is called confidence interval and the (predetermined) success rate of the procedure, $\text{Prob}([\underline{a}, \bar{a}] \ni a|a)$, is called confidence level. The statement $[\underline{a}, \bar{a}] \ni a$ maintains the unknown $[\underline{a}, \bar{a}]$ interval includes the known measurand value a , whereas $a \in [a_1, a_2]$ maintains the unknown measurand value a is included in the known $[a_1, a_2]$ interval.

In the same way as the measurement result, the confidence interval $[\underline{a}, \bar{a}]$ is picked at random from an interval set – the urn of the frequentist model – where the fraction of intervals containing the measurand value (the confidence level) is equal to the coverage probability. Therefore, $[\underline{a}, \bar{a}]$ is an estimate of the credible interval $[a_1, a_2]$ and the confidence level is a property of the estimator, non of the specific interval sampled. As Neyman repeatedly stated, the confidence level is only the probability that a future interval embeds the measurand.

The interval end-points are the solutions of

$$\text{Prob}([\underline{a}, \bar{a}] \ni a|a) = \alpha. \quad (3)$$

It must be noted that (3) is conditioned on the fixed measurand value a ; since it is unknown, (3) is meaningful only if the statistics used to calculate \underline{a} and \bar{a} is such that $\text{Prob}([\underline{a}, \bar{a}] \ni a|a)$ is independent of a . Afterwards, since (3) is independent of the actual measurand value, the probability statement about $[\underline{a}, \bar{a}]$ can be restated as one about the a value.

The Neyman procedure uses a pair of continuous and monotonic functions of the measurand, $x_1(a)$ and $x_2(a)$, so chosen as $\text{Prob}(x \in [x_1, x_2]|a) = \alpha$. That is,

$$F_x(x_2|a) - F_x(x_1|a) = \alpha, \quad (4)$$

where $F_x(\xi|a)$ is the cumulative distribution associated to $P_x(\xi|a)$. Provided the measured value x_0 is in the domain of both the inverse functions $x_2^{-1}(\xi)$ and $x_1^{-1}(\xi)$, it follows that $\underline{a} = x_2^{-1}(x_0)$ and $\bar{a} = x_1^{-1}(x_0)$ are the sought interval end-points. Equivalently, given the measured value x_0 , the interval end-points are the solution of

$$F_x(x_0|\underline{a}) - F_x(x_0|\bar{a}) = \alpha. \quad (5)$$

In the same way as a measurand estimate is not the measurand value, in general, the probability of $[\underline{a}, \bar{a}] \ni a$ – where $[\underline{a}, \bar{a}]$ is built by solving (5) and x_0 is a given measurement result – is not equal to the confidence-level [2]; that is, $\text{Prob}([\underline{a}, \bar{a}] \ni a|a) = \alpha$ but $\text{Prob}(a \in [\underline{a}, \bar{a}]|x_0) \neq \alpha$. Rather, $[\underline{a}, \bar{a}]$ is randomly sampled from a set of intervals where the fraction α of them embeds a . According to (3), this sample space is the set of the intervals built by solving (5), where x_0 are the results of repeated measurement of the same measurand.

Some remarks are needed. Firstly, (4) and (5) do not identify $[x_1, x_2]$ and $[\underline{a}, \bar{a}]$ uniquely; in order to have a single solution, additional constraints are necessary. Secondly, once a constraint has been chosen, there may exist measurement results wherefore (5) has no solution. Thirdly, when there are nuisance parameters, no general algorithm exists to build a confidence interval for the measurand value only.

2.1.1. Example: Gaussian measurement of a positive quantity. The figure 1 illustrates the Neyman procedure. The measured value x_0 of $a > 0$ is drawn from the normal distribution $P_x(\xi|a) = N(\xi; a, u^2)$ whose mean and variance are a and u^2 . For the sake of simplicity, the variance has been set to one, which corresponds to redefine the measurand and measured values as a/u and x_0/u . By setting $u^2 = 1$, the cumulative distribution is

$$F_x(\xi|a) = \frac{\text{erfc}[(a - \xi)/\sqrt{2}]}{2}, \quad (6)$$

where $\text{erfc}(x)$ is the complementary error function. In Fig. 1, the two lines, $x_1 = a - u$ and $x_2 = a + u$, are so chosen as $\text{Prob}(x \in [x_1, x_2]|a) = 0.68$. For the sake of simplicity, the numerical values are rounded to the second digit. Given the measured value x_0 , the solution of (5), where

$$F_x(x_0|\underline{a}) = 0.84 \quad (7a)$$

$$F_x(x_0|\bar{a}) = 0.16, \quad (7b)$$

is $[x_0 - u, x_0 + u]$; an example is the solid arrow in Fig. 1.

If $x_0/u < 1$, (5) has no solution satisfying (7a-b). A way to bypass this problem is to allow negative a values and to say that the confidence interval is partly non-physical. However, when $x_0/u < -1$, $[x_0 - u, x_0 + u]$ is entirely in the negative region and its coverage probability is null; that is, $\text{Prob}(a \in [x_0 - u, x_0 + u]|x_0) = 0$. This is not surprising; besides, a negative interval is not more unusual than a negative

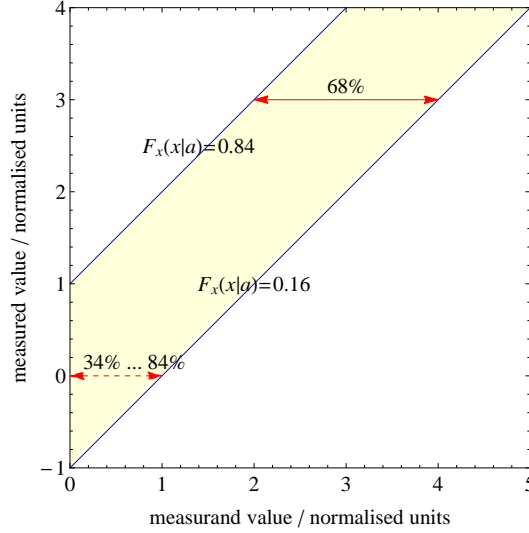


Figure 1. 16% and 84% quantiles of the results of an unbiased Gaussian measurement of a positive quantity. The solid arrow is the solution of (5) and (7a-b), where $x_0/u = 3$. The dashed arrow is the solution of (5) and (7b), where the confidence level is any value from 34% to 84% and $x_0/u = 0$.

datum and $[x_0 - u, x_0 + u]$ is one of the intervals of the Neyman's sample-space not including a . The paradox is solved by observing that it arises only because we know in advance that $a > 0$. In addition, it is not correct to identify the procedure confidence-level with the coverage probability of $[x_0 - u, x_0 + u]$. To say that $[x_0 - u, x_0 + u]$ is a 68% confidence interval means that $\text{Prob}([x - u, x + u] \ni a|a) = 0.68$, not that $\text{Prob}(a \in [x_0 - u, x_0 + u]|x_0) = 0.68$ [2]. The confidence level is conditioned to the measurand value, not to the measured value. This means that the probability of sampling a future interval such that $[x - u, x + u] \ni a$ is true, where x is unknown and a is known, is 0.68. But, once $x = x_0$ is on hand, the probability – updated by the information delivered by x_0 – of $a \in [x_0 - u, x_0 + u]$, where a is unknown and $[x_0 - u, x_0 + u]$ is known, might be different.

A second solution, proposed in [14, 20], is to exclude the $a < 0$ values. However, in this case, different confidence levels lead to the same interval. For instance, as shown by dashed arrow in Fig. 1, given (7b), $x_0/u = 0$, and any confidence level from 34% to 84%, the result is always the $[0, u]$ interval. A further solution is to switch between two-sided intervals and upper limits according to the measured value. For instance, if $x_0/u < 1$, to switch from (7a-b) to $\underline{a} = 0$ and $F_x(x_0|\underline{a}) = 0.32$. However, flip-flopping is inconsistent with a predetermined confidence level. A solution that uses the freedom to choose the $x_1(a)$ and $x_2(a)$ functions is given in [6]. The resulting intervals change continuously from upper limits to two-sided intervals as the measured value becomes more statistically significant.

2.2. Bayes: credible intervals

By definition, the probability of $a \in [a_1, a_2]$ is the integral of $P_a(\phi|x_0)$ between two given limits, a_1 and a_2 . Hence, the end points of credible intervals having a coverage

probability equal to α are the solutions of

$$F_a(a_2|x_0) - F_a(a_1|x_0) = \alpha, \quad (8)$$

where $F_a(\phi|x_0)$ is the cumulative distribution associated to $P_a(\phi|x_0)$. In the framework of a frequency-of-occurrence model of $\text{Prob}(a \in [a_1, a_2]|x_0)$, the sample space is the set of the a values consistent with the same measurement result and, consequently, with the same credible interval.

It must be noted that, to build credible intervals, the availability of a measurement result is not an essential ingredient. In fact, by resorting to the prior probability distribution $\pi(\phi)$, credible intervals can be built also if no measurement has been carried out. This emphasises again that a probability distribution is not an intrinsic quality of the measurand, but a way to encode our knowledge of its value.

When $\pi(\phi)$ is the uniform distribution and the probability density function of x owns the symmetry $P_x(\xi|\phi) = P_x(\phi|\xi)$, that is, it is invariant with respect to the replacement $\xi \rightleftharpoons \phi$, the Bayes theorem simplifies to $P_a(\phi|x_0) = P_x(\phi|x_0)$ and brings to light that the post-data probability density of the measurand values and the sampling distribution of the measurement results are the same function. If, in addition, $P_x(\xi|\phi)$ is a function of $|\xi - \phi|$ only, it can be proved that the Neyman and Bayesian procedures lead to the same interval. The occurrence of the interval identity – for instance, when $P_x(\xi|a)$ is the ubiquitous Gaussian distribution – causes misunderstandings. One may carry out a Neyman interval-estimation and use the result as if $\text{Prob}(a \in [\underline{a}, \bar{a}]|x_0) = \alpha$, which, in general, is not correct.

2.2.1. Example: Gaussian measurement of a positive quantity. Let us suppose again that the measured value x_0 of $a > 0$ is drawn from the normal distribution $P_x(\xi|a) = N(\xi; a, u^2)$. By setting again $u^2 = 1$, a uniform prior probability distribution of the a values must be updated into

$$P_a(\phi|x_0) = \frac{2 \exp[-(\phi - x_0)^2/2] \vartheta(\phi)}{\sqrt{2\pi} \text{erfc}(-x_0/\sqrt{2})}, \quad (9)$$

where $\text{erfc}(x)$ is the complementary error function and $\vartheta(\phi)$ is the Heaviside function. The relevant cumulative distribution is

$$F_a(\phi|x_0) = \frac{\text{erf}(x_0/\sqrt{2}) + \text{erf}[(\phi - x_0)/\sqrt{2}]}{\text{erfc}(-x_0/\sqrt{2})}, \quad (10)$$

where $\text{erf}(x)$ is the error function. Given the measured value x_0 , we will consider the intervals constrained by

$$F_a(a_1|x_0) = 0.16 \quad (11a)$$

$$F_a(a_2|x_0) = 0.84. \quad (11b)$$

After a measurement has been completed, the measurement result is a known quantity. Since we own only this unique result, it is not clear what population is to be used to imagine repeated measurements and to build a frequentist model of an unconditioned statement about the probability of the measurand to belong a given interval. In fact, there is a multiplicity of sample spaces to each of which we can regard the unknown $\{a, x\}$ repeated-measurement pairs as belonging, none having an objective reality and all being products of our subjective preference. However, (2) and (3) keep strictly to conditional statements so that the relevant frequentist models can be uniquely defined.

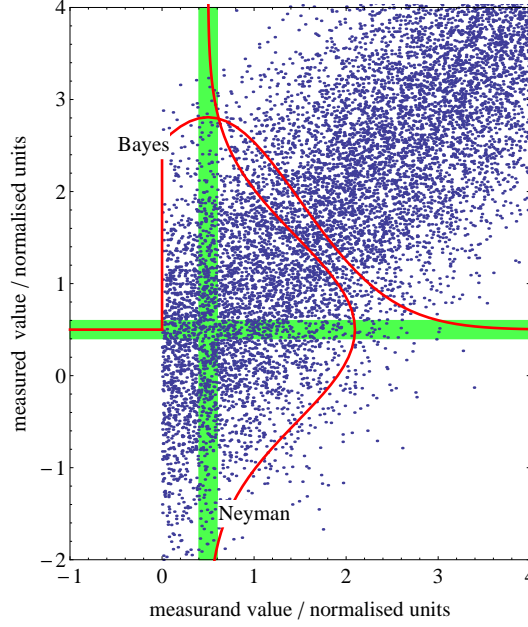


Figure 2. Scatter plot of the joint distribution of the $\{a_i, x_i\}$ pairs for a Gaussian measurement of a quantity uniformly distributed in the $[0, \infty]$ interval. The Neyman curve is the sampling distribution $N(\xi; a, u^2)$ of the measurement result for a measurand value $a/u = 0.5$. The Bayes curve is the post-data measurand distribution $P_a(\phi|x_0)$, given the measurement result $x_0/u = 0.5$. The green strips indicate the sample spaces – horizontal: $\{a_i, x = \text{const.}\}$, fixed measured value; vertical: $\{a = \text{const.}, x_i\}$, fixed measurand value – used to assess the success rates of the Bayes and Neyman solutions of the interval estimation problem.

The sample spaces of the frequency-of-occurrence models of $\text{Prob}([\bar{a}, \underline{a}] \ni a|a)$ and $\text{Prob}(a \in [a_1, a_2]|x_0)$ are shown in Fig. 2. The scatter plot shows the joint distribution $P_{x,a}(\xi, \phi) = N(\xi; \phi, u)\pi(\phi)$ of the measurand- and measured-value pairs $\{a_i, x_i\}$ for a Gaussian measurement of a quantity uniformly distributed in the $[0, \infty]$ interval. The pairs $\{a = \text{const.}, x_i\}$, having the same measurand value, make up the sample space of $\text{Prob}([\bar{a}, \underline{a}] \ni a|a)$ and will be used to assess the success rate of the Neyman intervals: in Fig. 2 they are in the vertical strip. The pairs $\{a_i, x = \text{const.}\}$, having the same measured value, make up the sample space of $\text{Prob}(a \in [a_1, a_2]|x_0)$ and will be used to assess the success rate the Bayesian intervals: in Fig. 2 they are in the horizontal strip.

3. Performance analysis

This section examines the performances of the Neyman and Bayesian procedures from a frequentist viewpoint. Monte Carlo simulations are used to calculate the success rates of confidence and credible intervals and to compare the results against the expected rates. The case studied is where the measurement of a positive quantity a gives a Gaussian datum having known variance u^2 .

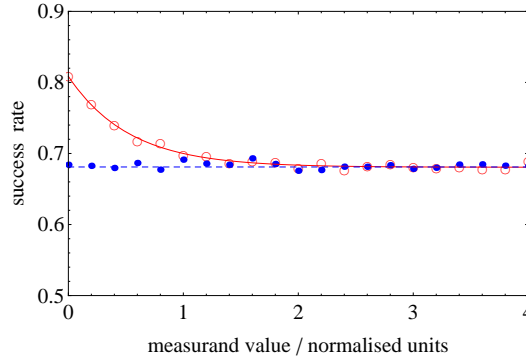


Figure 3. Gaussian measurements of a positive quantity: success rate of confidence intervals. Different intervals have been built from the results of repeated measurements of the same measurand. Dots: frequencies observed in Monte Carlo simulations; the horizontal line is the theoretically expected value $\text{Prob}([\underline{a}, \bar{a}] \ni a|a)$. The calculations have been carried out by setting (7a-b). Empty circles: the intervals entirely in the negative region have been rejected and measurements repeated; the solid line is a smoothed interpolation of the data.

3.1. Confidence intervals

According to (3), in a long series of repeated measurements of the same measurand, the fraction α of the different confidence intervals built by solving (5) contains the single a value. Since it is conditional on a fixed measurand value, to test numerically $\text{Prob}([\underline{a}, \bar{a}] \ni a|a) = \alpha$, the measurand value (say, $a_0 > 0$) must be fixed. Next, a number of measurement results are repeatedly sampled according to $P_x(\xi|a_0)$ and the relevant confidence intervals are built. Each trial involves determining if the interval contains the fixed measurand value. As shown by the horizontal line in Fig. 3, a Monte Carlo simulation, carried out by fixing $a/u > 0$ and building a confidence interval for each sample x_i , proves – not surprisingly – the effectiveness of the $[x_i - u, x_i + u]$ interval. The test has been carried out by setting the constraints (7a-b); negative intervals have been allowed.

The generation of non-physical intervals lying entirely in the negative region is crucial to comply with the stipulated confidence level. As shown in Fig. 3, if these intervals are rejected and the measurements repeated until a physically acceptable interval is observed, the confidence level of the procedure is higher than what stated in (3). Still worse, it is unpredictable, because it depends on the (unknown) measurand value.

3.2. Credible intervals

Since it must be conditional on a fixed measured value, a frequency-of-occurrence model of $\text{Prob}(a \in [a_1, a_2]|x_0)$ must rely on the $\{a_i, x = \text{const.}\}$ sample space; that is, on the set of measurand value consistent with a unique measured value and credible interval. The sampling from this space can be carried out as follows. Firstly, a measurand value (say, a_i) is sampled according to $\pi(\phi)$ – which encodes the pre-data information about a ; secondly, a measurement results is sampled according to $P_x(\xi|a_i)$; thirdly, if the measurement result is x_0 – in practice, to within some approximation – the a_i value is accepted; otherwise, it is rejected. In a long series of repetitions of this

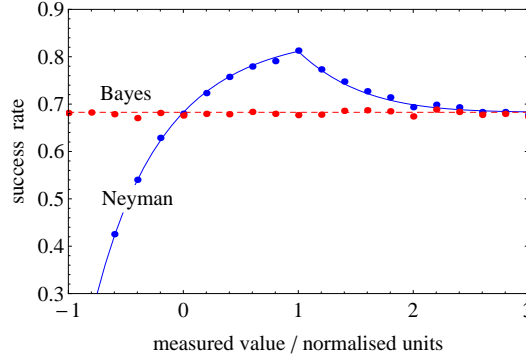


Figure 4. Gaussian measurements of a positive quantity: a single confidence interval and a single credible interval have been built for the same measured value repeatedly sampled according to different distributions $P_x(\xi|a)$, each distribution corresponding to a different measurand value. Neyman: success rates of confidence intervals; Bayes: success rates of credible intervals. Lines are the theoretically expected values $\text{Prob}(a \in [\underline{a}, \bar{a}]|x_0)$ and $\text{Prob}(a \in [a_1, a_2]|x_0)$; dots are the frequencies observed in Monte Carlo simulations. The calculations have been carried out by setting $(7a-b)$ and $(11a-b)$.

procedure, the fraction α of the accepted a_i values is expected to be inside the single credible interval built by solving (8).

In the case study here considered, the sampling from $\{a_i, x = \text{const.}\}$ and the assessment of the success rate of credible intervals can be carried out without assuming any prior distribution of the a values by the following numerical experiment. Firstly, a measurand value (say, a_0) is chosen in whichever way; next, a measurement results (say, x_i) is sampled according to $P_x(\xi|a_0)$. To have x_0 instead, the measurand value is shifted to $a_i = a_0 + x_0 - x_i$. If $a_i < 0$, the measurand value is rejected and the experiment is repeated. Otherwise, if $a_i > 0$, it is checked if a_i is in the fixed credible interval obtained by solving (8). It is worth noting that a uniform prior distribution of the measurand values emerges naturally from the model, without being predetermined. The test has been carried out by setting the constraints $(11a-b)$; not even saying, as the horizontal line in Fig. 4 shows, the observed success rate is 0.68.

3.3. Willink's paradox

The agreement of the long-run success rate of credible intervals with the predetermined coverage probability contradicts what observed by Willink [20] (Fig. 4 – solid line), which is reproduced by the solid line in Fig. 5. This figure shows that, when $a/u \lesssim 1$, the success rate of the Bayesian intervals disagrees with the expected value. The paradox is solved by observing that, in Fig. 5 and [20], the success rate is calculated conditionally on the measurand value; that is, by fixing the measurand value and by building a new credible interval for each different measured value. This is equivalent to calculate $\text{Prob}([a_1, a_2] \ni a|a)$. But Bayesian intervals are solutions of (2) and a frequency-of-occurrence model of $\text{Prob}(a \in [a_1, a_2]|x_0)$ must be conditional on the measured value; that is, it must rely on a fixed measurement result. Therefore, the paradox originates from the use of the sample space $\{a = \text{const.}, x_i\}$.

To investigate further the differences between (2) and (3), we calculated the probability – $\text{Prob}(a \in [\underline{a}, \bar{a}]|x_0) = F_a(\bar{a}|x_0) - F_a(\underline{a}|x_0)$, where $F_a(\phi|x_0)$ is given

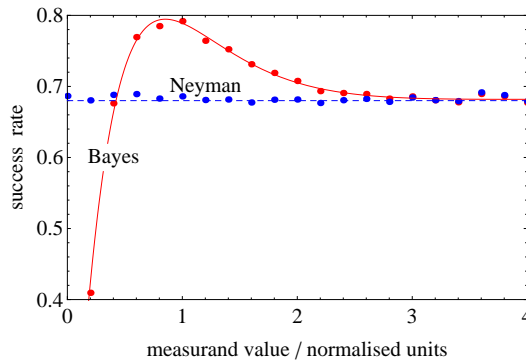


Figure 5. Gaussian measurements of a positive quantity. Dots: success rates of confidence – Neyman, the dashed line is the theoretically expected value $\text{Prob}([\underline{a}, \bar{a}] \ni a|a)$ – and credible – Bayes, the solid line is a smoothed interpolation of the data – intervals. The calculations have been carried out by setting (7a-b) and (11a-b), respectively. The sample space is $\{a = \text{const.}, x_i\}$: different credible- and confidence-intervals have been built from the results of repeated measurements of the same measurand.

by (10) – that the Neyman interval $[\underline{a}(x_0), \bar{a}(x_0)]$ built from the measured value x_0 embeds the (unknown) value of the measurand. The result is shown in Fig. 4. In addition, the success rate of $[\underline{a}, \bar{a}]$ has been calculated by a Monte Carlo simulation conditional on a fixed measured value; that is, by sampling from the $\{a_i, x = \text{const.}\}$ space, where x_i , which is known, is fixed and a , which is unknown, is random. This has been done by the same numerical experiment used to assess the success rate of credible intervals; as expected, Fig. 4 shows the poor performance of the Neyman procedure when tested in this way.

4. Conclusions

This paper examined interval estimation from both the Neyman and Bayesian viewpoints and investigated differences not always perceived. It demonstrated a frequentist model of the coverage probability of Bayesian intervals, where a single interval is built for the same measured value repeatedly sampled according to different distributions $P_x(\xi|a)$, each distribution corresponding to a different measurand value. No prior distribution of the measurand values has been explicitly assumed; rather, a uniform prior distribution emerges naturally from the model. Eventually, the paper proposed a solution to the paradoxical failure of the Bayesian intervals to pass a success-rate test based on repeated measurements of the same measurand and showed that an equivalent failure occurs when the Neyman intervals are tested against the same result repeatedly obtained by measuring different measurands.

The Neyman's view is the measurand value is not random; it is fixed and deterministic. Therefore, he discarded the specification of interval estimation given by (2) and turned to (3). This attitude and the requirement that $\text{Prob}([\underline{a}, \bar{a}] \ni a|a)$ is independent of the a value may lead to see the confidence level as the probability, $\text{Prob}(a \in [\underline{a}, \bar{a}]|x_0)$, of the measurand to be in a given $[\underline{a}, \bar{a}]$ interval, rather than what it is, namely the probability, $\text{Prob}([\underline{a}, \bar{a}] \ni a|a)$, of a future $[\underline{a}, \bar{a}]$ interval to encompass a given a value.

Both the Neyman and Bayesian approaches are correct, but confidence and credible intervals are solutions of different problems, namely (2) and (3). Hence, what is the best approach is an ill posed question. Whether to use one or the other to express the uncertainty of measurements depends on what problem we must solve and on decision theoretic considerations that are outside the scope of this paper. The following thoughts may supply some guidelines.

The construction of a generator of (random) intervals having a stipulated success rate of generating intervals including a fixed measurand, must rely on the Neyman procedure. For the Neyman's *practical statistician* in [2], the motivation of using confidence intervals lies in the customer satisfaction. If she sells confidence intervals, in the long run, she is sure that the fraction α of her customers had a correct statement. But, the probability of the measurand to be in any specific interval may be not equal to the success-rate: in his seminal paper, Neyman already stressed that the confidence level is not the probability that the measurand is in the calculated interval.

If we want to express the measurement uncertainty by stating the probability that the measurand value is within a stipulated interval, which statement is not implied by Neyman intervals, we must rely on the probability distributions $\pi(\phi)$, before the measurement, and $P_a(\phi|x_0)$, after the measurement. The need of a prior distribution is an unavoidable consequence of the product rule of probabilities that discourages the use of credible intervals, because of lack of objectivity. However, an objectivity request does not make the Bayes theorem to vanish; to calculate the probability that the measurand is embedded in a given interval without the use of a prior distribution is impossible.

In order to allow the decision makers to make the relevant inferences by combining the result with any other information they have, it is incumbent on metrologists to provide the probability distribution $P_x(\xi|a)$ or, at least, the variance of the population of the possible results. But, if we must come to a decision based on the measurand value (e.g., to choose a value of the Planck constant to redefine the mass unit) it is the posterior probability density $P_a(\phi|x_0)$ – hence, credible intervals – that we need in order to maximise the expected utility (e.g., the continuity of the kilogram realizations). In addition, to account for the model uncertainty, we need also the evidence of the measurement result $Z(x_0)$.

Acknowledgements

This work was jointly funded by the European Metrology Research Programme (EMRP) participating countries within the European Association of National Metrology Institutes (EURAMET) and the European Union.

References

- [1] Neyman J 1935 On the problem of confidence intervals *Ann. Math. Stat.* **6** 111-6
- [2] Neyman J 1937 Outline of a theory of statistical estimation based on the classical theory of probability *Philos. Trans. Roy. Soc. Ser. A* **236** 333-80
- [3] Jaynes E T 1976 Confidence intervals vs. Bayesian intervals in: *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science* vol. II 175-257 (Dordrecht, Holland: D. Reidel Publishing Company)
- [4] Sivia D S and Skilling J 2007 *Data Analysis: a Bayesian Tutorial* (Oxford: Oxford University Press)
- [5] Stein C 1959 An example of a wide discrepancy between fiducial and confidence intervals *Ann. Math. Stat.* **30** 877-80

- [6] Feldman G J and Cousins R D 1998 Unified approach to the classical statistical analysis of small signals *Phys. Rev. D* **57** 3873-89
- [7] D'Agostini G 2000 Confidence limits: What is the problem? Is there the solution *Workshop on confidence limits* eds. James F, Lyons L and Perrin Y (Genève: CERN)
- [8] Bukin A D 2003 A comparison of methods for confidence intervals *SLAC-R-703 Proceedings of PHYSTAT-2003: Statistical problems in particle physics, astrophysics and cosmology* eds. Lyons L, Mount R P and Reitmeyer R (Menlo Park: SLAC) 148-50
- [9] Lira I and Woeger W 2006 Comparison between the conventional and Bayesian approaches to evaluate measurement data *Metrologia* **43** S249-59
- [10] Hall B D 2008 Evaluating methods of calculating measurement uncertainty *Metrologia* **45** L5-8
- [11] Lira I 2008 On the long-run success rate of coverage intervals *Metrologia* **45** L21-3
- [12] Wang C M and Iyer H K 2009 Fiducial intervals for the magnitude of a complex-valued quantity *Metrologia* **46** 81-6
- [13] Possolo A, Toman B and Estler T 2009 Contribution to a conversation about the Supplement 1 to the GUM *Metrologia* **46** L1-7
- [14] Willink R 2010 On the validity of methods of uncertainty evaluation *Metrologia* **47** 80-9
- [15] Willink R 2010 Probability, belief and success rate: comments on 'On the meaning of coverage probabilities' *Metrologia* **47** 343-6
- [16] Attivissimo F, Giaquinto N and Savino M 2012 A Bayesian paradox and its impact on the GUM approach to uncertainty *Measurement* **45** 2194-202
- [17] Bergamaschi L, D'Agostino G, Giordani L, Mana G and Oddone M 2013 The detection of signals hidden in noise *Metrologia* **50** 269-76
- [18] Calónico D, Levi F, Lorini L and Mana G 2009 Bayesian inference of a negative quantity from positive measurement results *Metrologia* **46** 267-71
- [19] Calónico D, Levi F, Lorini L and Mana G 2009 Bayesian estimate of the zero-density frequency of a Cs fountain *Metrologia* **46** 629-36
- [20] Willink R 2010 Uncertainty in repeated measurement of a small non-negative quantity: explanation and discussion of Bayesian methodology *Accred. Qual. Assur.* **15** 181-8
- [21] Willink R 2010 Measurement of small quantities: further observations on Bayesian methodology *Accred. Qual. Assur.* **15** 521-7